Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences

Nikesh Garera, Chris Callison-Burch, David Yarowsky
Department of Computer Science, Johns Hopkins University
Baltimore MD, USA
{ngarera,ccb,yarowsky}@cs.jhu.edu

Abstract

This paper presents novel improvements to the induction of translation lexicons from monolingual corpora using multilingual dependency parses. We introduce a dependency-based context model that incorporates long-range dependencies, variable context sizes, and reordering. It provides a 16% relative improvement over the baseline approach that uses a fixed context window of adjacent words. Its Top 10 accuracy for noun translation is higher than that of a statistical translation model trained on a Spanish-English parallel corpus containing 100,000 sentence pairs. We generalize the evaluation to other word-types, and show that the performance can be increased to 18% relative by preserving part-of-speech equivalencies during translation.

1 Introduction

Recent trends in machine translation illustrate that highly accurate word and phrase translations can be learned automatically given enough parallel training data (Koehn et al., 2003; Chiang, 2007). However, large parallel corpora exist for only a small fraction of the world's languages, leading to a bottleneck for building translation systems in low-density languages such as Swahili, Uzbek or Punjabi. While parallel training data is uncommon for such languages, more readily available resources include small translation dictionaries, comparable corpora, and large amounts of monolingual data.

The marked difference in the availability of monolingual vs parallel corpora has led several researchers to develop methods for automatically learning bilingual lexicons, either by using monolingual corpora (Rapp, 1999; Koehn and Knight, 2002; Schafer and Yarowsky, 2002; Haghighi et al., 2008) or by exploiting the cross-language evidence of closely related "bridge" languages that have more resources (Mann and Yarowsky, 2001).

This paper investigates new ways of learning translations from monolingual corpora. We extend the Rapp (1999) model of context vector projection using a seed lexicon. It is based on the intuition that translations will have similar lexical context, even in unrelated corpora. For example, in order to translate the word "airplane", the algorithm builds a context vector which might contain terms such as "passengers", "runway", "airport", etc. and words in target language that have their translations (obtained via seed lexicon) in surrounding context can be considered as likely translations. We extend the basic approach by formulating a context model that uses dependency trees. The use of dependencies has the following advantages:

- Long distance dependencies allow associated words to be included in the context vector even if they fall outside of the fixed-window used in the baseline model.
- Using relationships like parent and child instead of absolute positions alleviates problems when projecting vectors between languages with different word orders.
- It achieves better performance than baseline context models across the board, and better performance than statistical translation models on Top-10 accuracy for noun translation when trained on identical data.

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 129–137,
Boulder, Colorado, June 2009.

We further show that an extension based on part-of-speech clustering can give similar accuracy gains for learning translations of all word-types, deepening the findings of previous literature which mainly focused on translating nouns (Rapp, 1999; Koehn and Knight, 2002; Haghighi et al., 2008).

2 Related Work

The literature on translation lexicon induction for low-density languages falls in to two broad categories: 1) Effectively utilizing similarity between languages by choosing a high-resource "bridge" language for translation (Mann and Yarowsky, 2001; Schafer and Yarowsky, 2002) and 2) Extracting noisy clues (such as similar context) from monolingual corpora with help of a seed lexicon (Rapp, 1999; Koehn and Knight, 2002; Schafer and Yarowsky, 2002, Haghighi et al., 2008). The latter category is more relevant to this work and is explained in detail below.

The idea of words with similar meaning having similar contexts in the same language comes from the Distributional Hypothesis (Harris, 1985) and Rapp (1999) was the first to propose using context of a given word as a clue to its translation. Given a German word with an unknown translation, a German context vector is constructed by counting its surrounding words in a monolingual German corpus. Using an incomplete bilingual dictionary, the counts of the German context words with known translations are projected onto an English vector. The projected vector for the German word is compared to the vectors constructed for all English words using a monolingual English corpus. The English words with the highest vector similarity are treated as translation candidates. The original work employed a relatively large bilingual dictionary containing approximately 16,000 words and tested only on a small collection of 100 manually selected nouns.

Koehn and Knight (2002) tested this idea on a larger test set consisting of the 1000 most frequent words from a German-English lexicon. They also incorporated clues such as frequency and orthographic similarity in addition to context. Schafer and Yarowsky, (2002) independently proposed using frequency, orthographic similarity and also showed improvements using temporal and word-burstiness similarity measures, in addition to context. Haghighi et al., (2008) made use of contextual and orthographic clues for learning a generative model from monolingual corpora and a seed lexicon.

All of the aforementioned work defines context similarity in terms of the adjacent words over a window of some arbitary size (usually 2 to 4 words), as initially proposed by Rapp (1999). We show that the model for surrounding context can be improved by using dependency information rather than strictly relying on adjacent words, based on the success of dependency trees for monolingual clustering and disambiguation tasks (Lin and Pantel, 2002; Pado and Lapata, 2007) and the recent developments in multilingual dependency parsing literature (Buchholz and Marsi, 2006; Nivre et al., 2007).

We further differentiate ourselves from previous work by conducting a second evaluation which examines the accuracy of translating all word types, rather than just nouns. While the straightforward application of context-based model gives a lower overall accuracy than nouns alone, we show how learning a mapping of part-of-speech tagsets between the source and target language can result in comparable performance to that of noun translation.

3 Translation by Context Vector Projection

This section details how translations are discovered from monolingual corpora through context vector projection. Section 3.1 defines alternative ways of modeling context vectors, and including baseline models and our dependency-based model.

The central idea of Rapp's method for learning translations is that of context vector projection and vector similarity. The goodness of semantic "fit" of candidate translations is measured as the vector similarity between two words. Those vectors are drawn from two different languages, so the vector for one word must first be projected onto the language space of the other. The algorithm for creating, projecting and comparing vectors is described below, and illustrated in Figure 1.

Algorithm:

1. Extract context vectors:

   Given a word in source language, say s_w, create a vector using the surrounding context words and call this reference source vector rs_{s_w} for

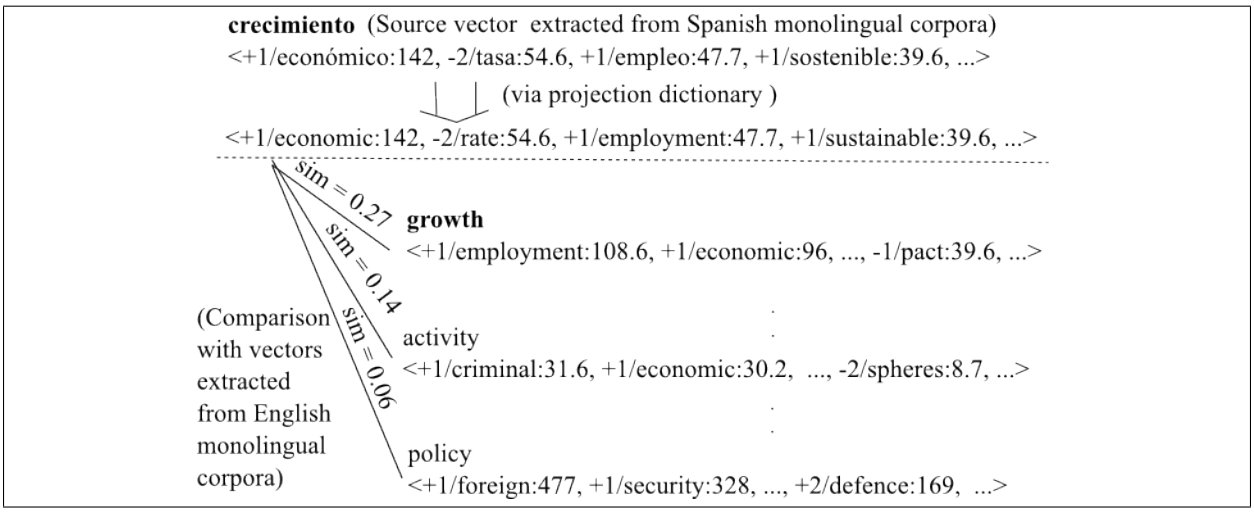

Figure 1: Illustration of (Rapp, 1999) model for translating spanish word "crecimiento (growth)" via dependency context vectors extracted from respective monolingual corpora as explained in Section 3.1.2

source word s_w. The actual composition of this vector varies depending on how the surrounding context is modeled. The context model is independent of the algorithm, and various models are explained in later sections.

2. Project reference source vector:

   Project all the source vector words contained in the projection dictionary onto the vector space for the target language, retaining the counts from source corpus. This vector now exists in the target language space and is called the reference target vector rt_{s_w}. This vector may be sparse, depending on how complete the bilingual dictionary is, because words without dictionary entires will receive zero counts in the reference target vector.

3. Rank candidates by vector similarity:

   For each word t_{w_i} in the target language a context vector is created using the target language monolingual corpora as in Step 1. Compute a similarity score between the context vector of $t_{w_i} = \langle c_{i1}, c_{i2}, ..., c_{in} \rangle$ and reference target vector $rt_{s_w} = \langle r_1, r_2, ..., r_n \rangle$. The word with the maximum similarity score $t^*_{w_i}$ is chosen as the candidate translation of s_w.

   The vector similarity can be computed in a number of ways. Our setup we used cosine similarity:

$$t^*_{w_i} = \text{argmax}_{t_{w_i}} \frac{c_{i1} \cdot r_1 + c_{i2} \cdot r_2 + ... + c_{in} \cdot r_n}{\sqrt{c_{i1}^2 + c_{i2}^2 + ... + c_{in}^2} \sqrt{r_1^2 + r_2^2 + ... + r_n^2}}$$

   Rapp (1999) used l1-norm metric after normalizing the vectors to unit length, Koehn and Knight (2002) used Spearman rank order correlation, and Schafer and Yarowsky (2002) use cosine similarity. We found that cosine similarity gave the best results in our experimental conditions. Other similarity measures may be used equally well.

3.1 Models of Context

We compared several context models. Empirical results for their ability to find accurate translations are given in Section 5.

3.1.1 Baseline model

In the baseline model, the context is computed using adjacent words as in (Rapp,1999; Koehn and Knight, 2002; Schafer and Yarowsky, 2002; Haghighi et al., 2008). Given a word in source language, say s_w, count all its immediate context words appearing in a window of four words. The counts are collected seperately for each position by keeping track of four seperate vectors for positions -2, -1, +1 and +2. Thus each vector is a sparse vector, having the # of dimensions as the size of source language vocabulary. Each dimension is also reweighted by multiplying the inverse document frequency (IDF)

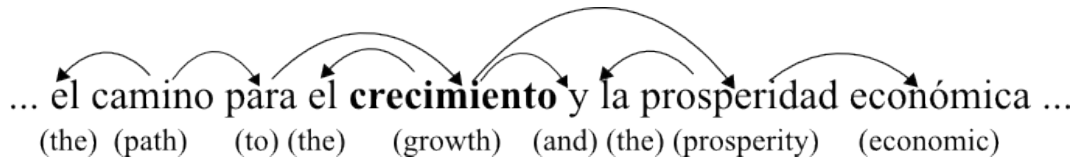

Figure 2: Illustration of using dependency trees to model richer contexts for projection

as in the standard TF.IDF weighting scheme[1]. These vectors are then concatenated into a single vector, having dimension four times the size of the vocabulary. This vector is called the reference source vector rs_{s_w} for source word s_w.

3.1.2 Modeling context using dependency trees

We use dependency parsing to extend the context model. Our context vectors use contexts derived from head-words linked by dependency trees instead of using the immediate adjacent lexical words. The use of dependency trees for modeling contexts has been shown to help in monolingual clustering tasks of finding words with similar meaning (Lin and Pantel, 2002) and we show how they can be effectively used for translation lexicon induction.

Position	Adjacent Context	Dependency Context
-2	para	camino
-1	el	para
+1	y	prosperidad, y, el
+2	la	económica

Table 1: Contrasting context words derived from the adjacent vs dependency models for the above example

The four vectors for positions -1, +1, -2 and +2 in the baseline model get mapped to immediate parent (-1), immediate child (+1), grandparent (-2) and grandchild (+2). An example of using the dependency tree context is shown in Figure 2, and the dependency context is shown in contrast with the adjacent context in Table 1, showing the selection of more salient words by using the dependencies.

Note that while we are limiting to four positions in the tree, it does not imply that only a maximum of four context words are selected since the word can have multiple immediate children depending upon the dependency parse of the sentence. Hence, this approach allows for a dynamic context size, with the number of context words varying with the number of children and parents at the two levels.

Another advantage of this method is that it alleviates the reordering problem as we use tree positions (consisting of head-words) as compared to the adjacent position in the baseline context model. For example, if the source spanish word to be translated was "prosperidad", then in the example shown in Figure 2, in case of adjacent context, the context word "económica" will show up in +1 position in Spanish and -1 position in English (as adjectives come before nouns in English) but in case of dependency context, the adjective will be the child of noun and hence will show up in +1 position in both languages. Thus, we do not need to use a bag of word model as in Section 3 in order to avoid learning the explicit mapping that adjectives and nouns in Spanish and English are reversed.

4 Experimental Design

For our initial set of experiments we compared several different vector-based context models:

- Adj_{bow} – A baseline model which used bag of words model with a fixed window of 4 words, two on either side of the word to be translated.
- Adj_{posn} – A second baseline that used a fixed window of 4 words but which took positional into account.
- Dep_{bow} – A dependency model which did not distinguish between grandparent, parent, child and grandparent relations, analogous to the bag of words model.
- Dep_{posn} – A dependency model which did include such relationships, and was analogous to the position-based baseline.
- $\text{Dep}_{\text{posn + rev}}$ – The above Dep_{posn} model applied in both directions (Spanish-to-English and English-to-Spanish) using their sum as the final translation score.

We contrasted the accuracy of the above methods, which use monolingual corpora, with a statistical

[1] In order to compute the IDF, while there were no clear document boundaries in our corpus, a virtual document boundary was created by binning after every 1000 words.

model trained on bilingual parallel corpora. We refer to that model as $\text{Moses}_{\text{en-es-100k}}$, because it was trained using the Moses toolkit (Koehn et al., 2007).

4.1 Training Data

All context models were trained on a Spanish corpus containing 100,000 sentences with 2.13 million words and an English corpus containing 100,000 sentences with 2.07 million words. The Spanish corpus was parsed using the MST dependency parser (McDonald et al., 2005) trained using dependency trees generated from the the English Penn Treebank (Marcus et al., 1993) and Spanish CoNLL-X data (Buchholz and Marsi, 2006).

So that we could directly compare against statistical translation models, our Spanish and English monolingual corpora were drawn from the Europarl parallel corpus (Koehn, 2005). The fact that our two monolingual corpora are taken from a parallel corpus ensures that the assumption that similar contexts are a good indicator of translation holds. This assumption underlies in all work of translation lexicon induction from comparable monolingual corpora, and here we strongly bias toward that assumption. Despite the bias, the comparison of different context models holds, since all models are trained on the same data.

4.2 Evaluation Criterion

The models were evaluated in terms of exact-match translation accuracy of the 1000 most frequent nouns in a English-Spanish dictionary. The accuracy was calculated by counting how many mappings exactly match one of the entries in the dictionary. This evaluation criterion is similar to the setup used by Koehn and Knight (2002). We compute the Top N accuracy in the standard way as the number of Spanish test words whose Top N English translation candidates contain a lexicon translation entry out of the total number of Spanish words that can be mapped correctly using the lexicon entries. Thus if "crecimiento, growth" is the correct mapping based on the lexicon entries, the translation for "crecimiento" will be counted as correct if "growth" occurs in the Top N English translation candidates for "crecimiento".

Note that the exact-match accuracy is a conservative estimate as it is possible that the algorithm may propose a reasonable translation for the given

camino			
Dep_{posn} Cntxt	Model	Adj_{bow} Cntxt	Model
way	0.124	intentions	0.22
solution	0.097	**way**	0.21
steps	0.094	idea	0.20
path	0.093	thing	0.20
debate	0.085	faith	0.18
account	0.082	**steps**	0.17
means	0.080	example	0.17
work	0.079	news	0.16
approach	0.074	work	0.16
issue	0.073	attitude	0.15

Table 2: Top 10 translation candidates for the spanish word "camino (way)" for the best adjacent context model (Adj_{bow}) and best dependency context model (Dep_{posn}). The bold English terms show the acceptable translations.

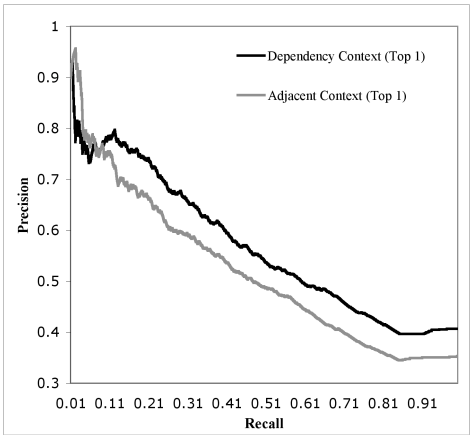

Figure 3: Precision/Recall curve showing superior performance of dependency context model as compared to adjacent context at different recall points. Precision is the fraction of tested Spanish words with Top 1 translation correct and Recall is fraction of the 1000 Spanish words tested upon.

Spanish word but is marked incorrect if it does not exist in the lexicon. Because it would be intractable to compare each projected vector against the vectors for all possible English words, we limited ourselves to comparing the projected vector from each Spanish word against the vectors for the 1000 most frequent English nouns, following along the lines of previous work (Koehn and Knight, 2002; Haghighi et al., 2008).

5 Results

Table 3 gives the Top 1 and Top 10 accuracy for each of the models on their ability to translate Spanish nouns into English. Examples of the top 10 translations using the best performing baseline and dependency-based models are shown in Table 2. The baseline models Adj_{posn} and Adj_{bow} differ in that the

Model	$\text{Acc}_{\text{Top 1}}$	$\text{Acc}_{\text{Top 10}}$
Adj_{bow}	35.3%	59.8%
Adj_{posn}	20.9%	46.9%
Dep_{bow}	41.0%	62.0%
Dep_{posn}	41.0%	64.1%
$\text{Dep}_{\text{posn + rev}}$	42.9%	**65.5%**
$\text{Moses}_{\text{en-es-100k}}$	**56.4%**	62.7%

Table 3: Performance of various context-based models learned from monolingual corpora and phrase-table learned from parallel corpora on Noun translation.

latter disregards the position information in the context vector and simply uses a bag of words instead. Table 3 shows that Adj_{bow} gains using this simplification. A bag of words vector approach pools counts together, which helps to reduce data sparsity. In the position based model the vector is four times as long. Additionally, the bag of words model can help when there is local re-ordering between the two languages. For instance, Spanish adjectives often follow nouns whereas in English the the ordering is reversed. Thus, one can either learn position mappings, that is, position +1 for adjectives in Spanish is the same as position -1 in English or just add the the word counts from different positions into one common vector as considered in the bag of words approach.

Using dependency trees also alleviates the problem of position mapping between source and target language. Table 3 shows the performance using the dependency based models outperforms the baseline models substantially. Comparing Dep_{bow} to Dep_{posn} shows that ignoring the tree depth and treating it as a bag of words does not increase the performance. This contrasts with the baseline models. The dependency positions account for re-ordering automatically. The precision-recall curve in Figure 3 shows that the dependency-based context performs better than adjancet context at almost all recall levels.

The $\text{Moses}_{\text{en-es-100k}}$ model shows the performance of the statistical translation model trained on a bilingual parallel corpus. While the system performs best in Top 1 accuracy, the dependency context-based model that ignores the sentence alignments surprisingly performs better in case of Top 10 accuracy, showing substantial promise.

While computing the accuracy using the phrase-table learned from parallel corpora ($\text{Moses}_{\text{en-es-100k}}$), the translation probabilities from both directions ($p(es|en)$ and $p(en|es)$) were used to rank the candidates. We also apply the monolingual context-based model in the reverse direction (from English to Spanish) and the row with label $\text{Dep}_{\text{posn + rev}}$ in Table 3 shows further gains using both directions.

Spanish	English	Sim Score	Is present in lexicon
señores	gentlemen	0.99	NO
xenofobia	xenophobia	0.87	YES
diversidad	diversity	0.73	YES
chipre	cyprus	0.66	YES
mujeres	women	0.65	YES
alemania	germany	0.65	YES
explotación	exploitation	0.63	YES
hombres	men	0.62	YES
república	republic	0.60	YES
racismo	racism	0.59	YES
comercio	commerce	0.58	YES
continente	continent	0.53	YES
gobierno	government	0.52	YES
israel	israel	0.52	YES
francia	france	0.52	YES
fundamento	certainty	0.51	NO
suecia	sweden	0.50	YES
tráfico	space	0.49	NO
televisión	tv	0.48	YES
francesa	portuguese	0.48	NO

Table 4: List of 20 most confident mappings using the dependency context based model for noun translation. Note that although the first mapping is the correct one, it was not present in the lexicon used for evaluation and hence is marked as incorrect.

6 Further Extensions: Generalizing to other word types via tagset mapping

Most of the previous literature on this problem focuses on evaluating on nouns (Rapp, 1999; Koehn and Knight 2002; Haghighi et al., 2008). However the vector projection approach is general, and should be applicable to other word-types as well. We evaluated the models with new test set containing 1000 most frequent words (not just nouns) in the English-Spanish lexicon.

We used the dependency-based context model to create translations for this new set. The row labeled Dep_{posn} in Table 5 shows that the accuracy on this set is lower when compared to evaluating only on nouns. The main reason for lower accuracy is that closed class words are often the most frequent and tend to have a wide range of contexts resulting in reasonable translation for most words include open class words via the context model. For instance, the English preposition "to" appears as the most confident translation for 147 out of the 1000 Spanish test

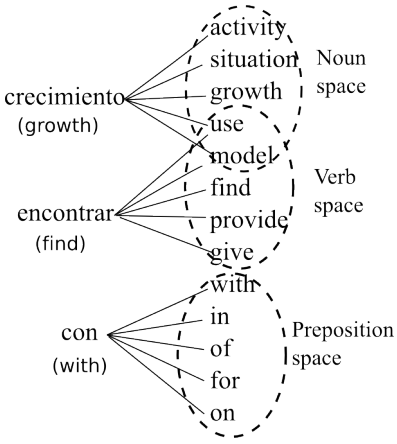

Figure 4: Illustration of using part-of-speech tag mapping to restrict candidate space of translations

words and in none (rightly so) after restricting the translations by part-of-speech categories.

This problem can be greatly reduced by making use of the intuition that part-of-speech is often preserved in translation, thus the space of possible candidate translation can be largely reduced based on the part-of-speech restrictions. For example, a noun in source language will usually be translated as noun in target language, determiner will be translated as determiner and so on. This idea is more clearly illustrated in in Figure 4. We do not impose a hard restriction but rather compute a ranking based on the conditional probability of candidate translation's part-of-speech tag given source word's tag.

An interesting problem in using part-of-speech restrictions is that corpora in different languages have been tagged using widely different tagsets and the following subsection explains this problem in detail:

6.1 Mapping Part-of-Speech tagsets in different languages

The English tagset was derived from the Penn treebank consisting of 53 tags (including punctuation markers) and the Spanish tagset was derived from the Cast3LB dataset consisting of 57 tags but there is a large difference in the morphological and syntactic features marked by the tagset. For example, the Spanish tagset as different tags for masculine and feminine nouns and also has a different tag for coordinated nouns, all of which need to be mapped to the singular or plural noun category available in English tagset. Figure 5 shows an illustration of the mapping problem between the Spanish and English POS tags.

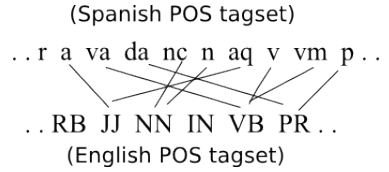

Figure 5: Illustration of mapping Spanish part-of-speech tagset to English tagset. The tagsets vary greatly in notation and the morphological/syntactic constituents represented and need to be mapped first, using the algorithm described in Section 6.1.

We now describe an empirical approach for learning the mapping between tagsets using the English-Spanish projection dictionary used in the monolingual context-based models for translation. Given a small English-Spanish bilingual dictionary and a n-best list of part-of-speech tags for each word in the dictionary[2], we compute conditional probability of translating a source word with pos tag s_{pos_i} to a target with pos tag t_{pos_j} as follows:

$$p(t_{pos_j}|s_{pos_i}) = \frac{c(s_{pos_i}, t_{pos_j})}{c(s_{pos_i})} = \frac{\sum_{s_w \in S, t_w \in T} p(s_{pos_i}|s_w) \cdot p(t_{pos_j}|t_w) \cdot I_{dict}(s_w, t_w)}{\sum_{s_w \in S} p(s_{pos_i}|s_w)}$$

where

- S and T are the source and target vocabulary in the seed dictionary, with s_w and t_w being any of the words in the respective sets.
- $p(s_{pos_i}|s_w), p(t_{pos_j}|t_w)$ are obtained using relative frequencies in a part-of-speech tagged corpus in the source/target languages respectively, and are used as soft counts.
- $I_{dict}(s_w, t_w)$ is the indicator function with value 1 if the pair (s_w, t_w) occurs in the seed dictionary and 0 otherwise.

In essence, the mapping between tagsets is learned using the known translations from a small dictionary.

Given a source word s_w to translate, its most likely tag s^*_{pos}, and the most likely mapping of this tag into English t^*_{pos} computed as above, the translation candidates with part-of-speech tag t^*_{pos} are considered for comparison with vector similarity and

[2]The n-best part-of-speech tag list for any word in the dictionary was derived using the relative frequencies in a part-of-speech annotated corpora in the respective languages

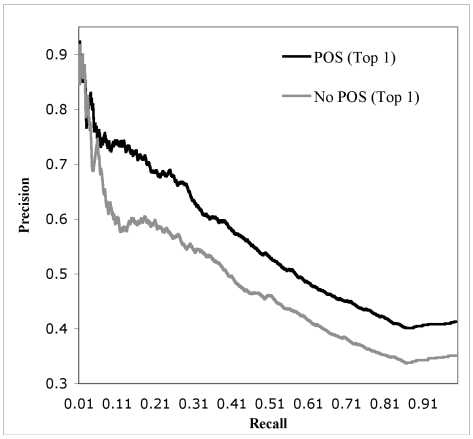

Figure 6: Precision/Recall curve showing superior performance of using part-of-speech equivalences for translating all word-types. Precision is the fraction of tested Spanish words with Top 1 translation correct and Recall is fraction of the 1000 Spanish words tested upon.

the other candidates with $t_{pos_j} \neq t^*_{pos}$ are discarded from the candidate space. Figure 4 shows an example of restricting the candidate space using POS tags.

Model	$\text{Acc}_{\text{Top 1}}$	$\text{Acc}_{\text{Top 10}}$
Dep_{posn}	35.1%	62.9%
+ POS	**41.3%**	**66.4%**

Table 5: Performance of dependency context-based model along with addition of part-of-speech mapping model on translating all word-types.

The row labeled $+POS$ in Table 5 shows the part-of-speech tags provides substantial gain as compared to direct application of dependency context-based model and is also comparable to the accuracy obtained evaluating just on nouns in Table 3.

7 Conclusion

This paper presents a novel contribution to the standard context models used when learning translation lexicons from monolingual corpora by vector projection. We show that using contexts based on dependency parses can provide more salient contexts, allow for dynamic context size, and account for word reordering in the source and target language. An exact-match evaluation shows 16% relative improvement by using a dependency-based context model over the standard approach. Furthermore, we show that our model, which is trained only on monolingual corpora, outperforms the standard statistical MT approach to learning phrase tables when trained on the same amount of sentence-aligned parallel corpora, when evaluated on Top 10 accuracy.

Spanish	English	Sim Score	Is present in lexicon
señores	gentlemen	0.99	NO
chipre	cyprus	0.66	YES
mujeres	women	0.65	YES
alemania	germany	0.65	YES
hombres	men	0.62	YES
expresar	express	0.60	YES
racismo	racism	0.59	YES
interior	internal	0.55	YES
gobierno	government	0.52	YES
francia	france	0.52	YES
cultural	cultural	0.51	YES
suecia	sweden	0.50	YES
fundamento	basis	0.48	YES
francesa	french	0.48	YES
entre	between	0.47	YES
origen	origin	0.46	YES
tráfico	traffic	0.45	YES
de	of	0.44	YES
social	social	0.43	YES
ruego	thank	0.43	NO

Table 6: List of 20 most confident mappings using the dependency context with the part-of-speech mapping model translating all word-types. Note that although the second best mapping in Table4 for noun-translation is for xenofobia with score 0.87, xenofobia is not among the 1000 most frequent words (of all word-types) and thus is not in this test set.

As a second contribution, we go beyond previous literature which evaluated only on nouns. We showed how preserving a word's part-of-speech in translation can improve performance. We further proposed a solution to an interesting sub-problem encountered on the way. Since part-of-speeech tagsets are not identical across two languages, we propose a way of learning their mapping automatically. Restricting candidate space based on this learned tagset mapping resulted in 18% improvement over the direct application of context-based model to all word-types.

Dependency trees help improve the context for translation substantially and their use opens up the question of how the context can be enriched further making use of the hidden structure that may provide clues for a word's translation. We also believe that the problem of learning the mapping between tagsets in two different languages can be used in general for other NLP tasks making use of projection of words and its morphological/syntactic properties between languages.

References

S. Buchholz and E. Marsi. 2006. Conll-X shared task on multilingual dependency parsing. *Proceedings of CoNLL*, pages 189–210.

Y. Cao and H. Li. 2002. Base Noun Phrase translation using web data and the EM algorithm. *Proceedings of COLING-Volume 1*, pages 1–7.

D. Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

P. Fung and L.Y. Yee. 1998. An IR Approach for Translating New Words from Nonparallel, Comparable Texts. *Proceedings of ACL*, 36:414–420.

A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. *Proceedings of ACL-HLT*, pages 771–779.

Z. Harris. 1985. Distributional structure. *Katz, J. J. (ed.), The Philosophy of Linguistics*, pages 26–47.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL-HLT*, pages 48–54.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. *Proceedings of ACL, companian volume*, pages 177–180.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit X*.

D. Lin and P. Pantel. 2002. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(04):343–360.

G.S. Mann and D. Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. *Proceedings of NAACL*, pages 151–158.

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. *Proceedings of EMNLP-HLT*, pages 523–530.

J. Nivre, J. Hall, S. Kubler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The conll 2007 shared task on dependency parsing. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 915–932.

S. Pado and M. Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2):161–199.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. *Proceedings of ACL*, pages 519–526.

C. Schafer and D. Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. *Proceedings of COLING*, pages 1–7.